VOICE IN OFFICE AUTOMATION - IMPLEMENTATION OF A VOICE RESPONSE SYSTEM APPLICATION

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of

M.Tech.

by

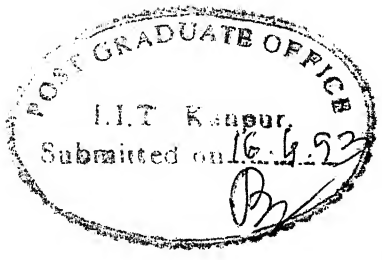
Rajesh S. Bhajikhaye

to the

DEPARTMENT OF INDUSTRIAL & MANAGEMENT ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

APRIL, 1993



CERTIFICATE

This is to certify that the present work on *"Voice In Office Automation - Implementation Of A Voice Response System Application"*, by Rajesh S. Bhajikhaye has been carried out under my supervision and has not been submitted elsewhere for the award of a degree.

(S. SADAGOPAN)

Professor and Head

April, 1993

Industrial & Management Engineering

Indian Institute of Technology

Kanpur-208016

ACKNOWLEDGEMENT

With deep sense of respect and gratitude, I thank Dr. S. Sadagopan for his valuable guidance. His different experiences, he shared with us will be very helpful to me in my life.

I am also thankful to all members of the IME family. Shiva, Ajay Misra, Raj, Alok, Shankar, Rakesh, Rajan accompanied me in the lab. I am thankful to these people as well as Bimal Modi and Banerjee who, instead of their busy schedule spent time to help me. Srikaanth, Mathews, Piyush, Saibal and Prabhakar have also provided me a very good company.

I am also thankful to all members of the AGAIK, who made my stay at IIT very cheerful. My friend Ravindra Prasad's discussions have been very fruitful to me, I am thankful to him.

RAJESH S. BHAJIKHAYE

CONTENTS

LIST OF FIGURES

ABSTRACT

In the present work an attempt has been made to develop an application of voice response system, "*A Railway Announcement System at Kanpur Railway Station*", to demonstrate the use of voice in Office Automation. The benefits of the application for an operator as well as for a listener are also mentioned.

The application has been developed to run on DOS, since the application makes the use of Speech Thing System (from COVOX Co.) which allows Text-To-Speech conversion. The source code has been written in C, using Borland C++. The application proves to be useful, though implemented on an experimental basis.

In addition to developing of an application, the different techniques and uses of voice are explained.

CHAPTER 1

INTRODUCTION

1.1 OFFICE AUTOMATION

1.1.1 OFFICE AS INFORMATION PROCESSING CENTER

" Better management results from better decisions. Better decisions result from better information. "

Information is an essential ingredient in Decision Making. Decision Making forms an integral part of the management. Thus information plays a significant role in management.

In the past, information was viewed as a 'Paper Dragon' that could potentially strangle the firm and prevent it from doing real work. By 1960 the concept of information has changed. Today information is viewed as a resource by organizations. Information became a tool to aid management in providing a fine tuned, special purpose and customized control over the organization. Today leading companies and organizations are using information technology as a competitive tool to develop new products and services, forge new relationships with suppliers, edge out competitors and radically change their internal operations and organizations. In other words we are witnessing an information driven management of organizations.

Earlier, information used to be achieved by plain observation or through oral communication. But with passage of time, quantum of information has significantly increased. Information must be timely, accurate and relevant since that alone ensures effective Decision Making. Information plays a key role in the affairs of organization and acts as a mechanism which controls and coordinates functioning of an enterprise.

To fulfill all these needs of an information in modern times, the information is required to be gathered by sophisticated computers.

Generally in offices four modes are considered for conveyance of an information. These modes are

1) Data - It is the information in non sentence form,

2) Text - words in sentence form,

3) Graphics - pictures and

4) Audio - sound.

An office deals with and "manufactures" only one commodity, i.e., information [2]. Even if it is obvious what an office is, there are a number of definitions about an office. Jarret [12] defines " An office is an information processing center : information flows in, it gets stored, and retrieved, and amended, and reformatted with other information, it gets distributed, it gets used. " Hirchheim [10] summarizes various offices by defining " offices can be thought of as centers of organizational information handling and processing. An Office

* receives information,
* records and stores information,
* structures information,
* processes information,
* provides access to information. "

Doswell [4] describes offices as consisting of components : machines, procedures and people. These components are concerned with

* the production of numbers and words,
* the storage and retrieval of numbers and words,
* the communications of numbers and words.

From the foregoing definitions it emerges that an office is essentially an information processing center with its basic elements : people, machines, paper, records, documents, files and procedures. It is concerned with inflow, storage, sifting, reformatting and dissemination of information.

Every organization sets a set of objectives which it tries to achieve. All activities performed in the office are mere mechanisms to achieve the objectives. Office functions can be described by four categories of office operations, viz : Document Preparation, Message Distribution, Personal Information Management and Information access [22].

The Office system researchers have listed the following types of office work : Document creation and preparation (handwriting, typing, dictating, drawing, diagrams, printing), Information/Document storage, Information/Document retrieval, Communication, Meetings, Travel to Meetings, Reading mail/documents, Thinking and Decision Making, Information/Document dissemination.

All these are clearly information handling activities and emphasize that office is primarily an information processing center.

1.1.2 **OFFICE AUTOMATION**

In a broad sense, Office Automation is the use of appropriate technology to help people manage information. Office Automation is the linking of multiple components or elements in order to process and channel information with a minimum of human intervention.

Hirchheim [10] has collected a number of definitions of office automation. Some of these are as folllows :

As defined by Olson and Lumas , Office Automation refers to the use of integrated computer and communication systems to support administrative procedures in an office environment.

According to Ellis and Nutt, an automated office system attempts to perform the functions of the ordinary office by means of a computer system.

Hammer and Sirbu consider Office Automation as the utilization of the technology to improve the realization of office functions.

Summarizing, we can say that Office Automation is use of information technology for Generation, Storage and Retrieval, Processing and Communication of information for improving the effectiveness of office, which in turn will help to realize the objectives or business functions of the organization in an efficient and competitive manner.

Office Automation incorporates technology to serve rather than be served by the people, for, people are more valuable than the equipment. System planners are now incorporating principles of human factors engineering into the design of hardware as well as software, resulting in user-friendly systems.

Except for data processing and photocopying, there was little concentration on office technology prior to the 1980s [3]. Now, however, technology is flooding in the office environment, and with it has come a marked improvement in the design and implementation of tools for information management.

Good office systems are modular and highly flexible, providing users with as much freedom of choice as possible. They offer versatility and transparency of time and place, that is the local time of the day and location of people has little or no impact, even if the people involved are in different time zones or different geographic locations. Personal computers can enable employees to access electronic files and documents in

seconds, at any time, from any place. Users of terminals or personal computers can easily and rapidly communicate with a number of people in the same building or on the opposite sides of the earth. Because of office automation, one person can take part in the meetings in different cities in one day without leaving his office.

Virtually the same functional elements tend to recur in the system design of the office automation plans, embracing four modes of information conveyance, i.e. data, text, graphics and audio. These functions are :

1) **Text Management** : Text management means the capturing, manipulation, output, and storage of words and sometimes graphics images. Text management includes electromechanical and electronic typewriters, computer text editors, word processing systems operated on computers. Power and usefulness of word processors far exceed those of typewriters. They provide faster keyboarding and easier formatting of complex documents or text layouts. They also eliminate the need to print information on paper until every portion of it is correct. Now-a-days word processing systems equipped with communication capabilities are possessing an immense versatility.

2) **Electronic filing and Data Base Management** : Electronic filing and retrieving is a rapid and accurate means of accessing information. It is an important component of an Office automation system as Word processors, computer based message systems, electronic calendars and reminders, micrographic units must file (store)and retrieve files of information. Magnetic disks and floppy diskettes are the most important media of storage of the information, but optical disks are making rapid inroads because of their immense, yet, compact storage capabilities.

Database Management Systems contain everything necessary to define, create, maintain, and modify the database both in form and content; and it provides a means of inquiry and report generation. A Query language lets the user quickly write a program to perform queries and to print reports in just minutes.

3) **Electronic mail System** : An electronic mail system (E-mail) is a point to point conveyer of audio, data/text, graphic, modes of information. It can be anything from two telephones to the most advanced computer messaging systems. Today electronic mail systems are either computerized or non-computerized. Many computerized systems consist of terminals organized around network. Electronic mail system provides many advantages like faster delivery of information, less paperwork and photocopying, reduced mailing expenses, geographic independence, improved access to personnel.

4) **Teleconferencing** : One of the most powerful inventions of office automation is teleconferncing i.e., meetings without meeting because it provides a way to reduce the vast amount of time as well as the vast amount of money on meetings. Teleconferencing allows people, either individual or groups at distant locations to meet by means of telecommunications. Teleconferencing ranges from a simple long-distance telephone call to a complex electronic integration of audio and video and text. It often lets you have a meeting when time constraints or some situations threaten to delay or cancel the plans. Some of the benefits of the teleconferncing are reduced time and money of travel, ability to attend several meetings at diverse locations in a single day, quicker solutions to the problems and early implementation of the results of meetings.

5) **Micrographics** : Micrographics is the capture, retrieval and display of miniaturized, high resolution photographic images containing either textual or graphic information. The medium of micrographics is usually film, sometimes paper. Today micrographics is also a valuable part of office automation. It provides for quick and inexpensive duplication of images to be distributed to a group, for easy viewing by projection or computerized display, and for re-creation of hard copies of microforms. Micrographics offers compact maintenance of active files and archival storage. With its array of film types, techniques, and retrieval options, micrographics affords the end user many benefits like economy of document and film creation economy and rapidity of duplication, compact storage, high speed of retrieval, portability. Along with these benefits micrographics offers a few limitations like no space provided for annotation, need to access to viewer or a terminal, appropriate system for frequent updating, serial accessibility of the film.

1.1.3 BENEFITS OF OFFICE AUTOMATION

Manager's primary aim in business is to make profit. This is possible by increasing the productivity (productivity is the ratio of output versus input). Office automation enables him to increase the productivity.

When the office automation is properly implemented, we can list down its benefits as follows :

1) Increase productivity

2) Improve accuracy

3) Speed up throughput

4) Speed up turnaround

5) Gain competitive edge

6) Improve timeliness of information

7) Improve decision making

8) Conserve natural resources

9) Increase scope of control

10) Enhance individual and organizational flexibility

11) Make information portable

12) Decrease expenses

13) Reduce capital investment in structures

14) Reduce payroll costs

15) Optimize staffing

16) Enhance human capabilities

17) Conserve human resources

18) Compensate for human shortage

1.2 VOICE IN OFFICE AUTOMATION

Voice is the main form of communication in the office. Various studies have indicated that managers feel more comfortable in communicating through speech, rather than by writing. Managers are known to have a certain aversion to keyboard either for data entry or data retrieval. In the office, manager's most important areas are communication, meetings and decision making activities. Managers can expect to improve their productivity only if their areas are addressed in the office automation.

Many current office systems have made advances in user friendliness with the use of mouse and icons in Windows environment. The workstation screen is used to represent desktop, with small icons used to depict objects that may be found in the office. To begin a task, the user opens one of the objects on the desktop. The desired object is selected by using

a mouse to point to the icon on the screen. When the object is opened, a window on the screen is created to display its contents. Such interfaces offer many potential improvements over conventional systems. However, they still limit user to communicating with the system via the screen and keyboard (and mouse). One way to overcome this limitation is to integrate voice communication into office system.

There are many reasons for the drive to use the voice in the office :

It is possible to communicate emotional cues such as excitement, anger, satisfaction or doubt far more easily when using voice than with other media.

Voice is the most natural form of communication. Nearly everyone is comfortable in speaking, but many feel limitations in using a keyboard.

Voice is a speedier form of input, the people can dictate vocally up to six times faster than they can write. For many office workers, who are not accomplished typists, speech is faster than keyboard entry.

1.2.1 VOICE WITH COMPUTERS

Although voice may be the most natural way for people to communicate with computers, voice processing has played little role in most of the office systems to date. However, the falling costs of processing power and computer storage, linked to the rapid improvement in speech compression techniques, have resulted in the announcement of many new types of voice products which use computer technology. Following are the major voice related technologies which are explained in the chapter 3.

1) **Voice Response** is the process of generating human like voice from a machine. Systems have emerged which are capable of converting electronic text to synthesized speech.

2) **Voice Recognition** involves the inputting of data to a computer using a human voice rather than using a keyboard or any other means. Currently the technology is restricted to the issuing of commands to the system, but it is likely to develop in the area of creation of electronic documents by voice input.

3) **Voice Annotation** is the facility to attach a vocal comment to an electronic document, in much the same way that a paper document may be annotated in a handwritten fashion.

4) **Voice Messaging Systems** extend the capability of text-based electronic mail facilities to voice messaging. Typical features include the ability to leave voice messages or access voice messages from, a telephone handset, to store or modify the messages, to deliver them at a specified time and to broadcast them to an entire distribution list. Such systems are also sometimes referred to as Voice Store and Forward Systems.

In the office systems these voice technologies can be used for communication between user and the office system as the user interface to system as well as the communication between people along with the various categories of office information (i.e. text, data, graphics).

1.2.2 BENEFITS OF VOICE IN OFFICE AUTOMATION

Voice is desirable in office systems because

1) Voice is efficient. Dictation studies have shown that people can dictate vocally up to six times faster than they can write; for the many office workers who are not accomplished typists, speech is the faster than keyboard entry.

2) Voice is effective. Emotions such as excitement, hesitation, anger and satisfaction are most easily communicated by voice than any other media. These elements are very important in office communications.

3) Voice is natural. Nearly everyone is comfortable in speaking, but fewer feel comfortable using a keyboard.

Disadvantages of the voice are, it is slow to listen to, and is not easy to access randomly. Also people may feel awkward talking to a machine, due to lack of feedback that is normally received when speaking to other people. These problems can be overcome by using the voice media in conjunction with workstation keyboard and display. A voice editor program can be used to enter and revise voice data, since the voice editor displays a visual representation of the user's voice on the workstation screen. This may provide the feedback necessary to help overcome any reluctance to talk to a machine. The visual representation, combined with playback speed control, instantaneous access to any portion of the recorded voice and the ability to edit sections of the voice, may help overcome many disadvantages associated with voice playback.

1.3 ORGANIZATION OF THE THESIS

In this thesis an attempt has been made to discuss the use of voice in office automation. By voice we mean the voice with computers. In this thesis, the different technologies of voice with computers and different methods to represent voice in computer are also explained in brief. And to demonstrate the use of voice in office automation, an application has been developed.

In chapter 2, first the speech production system of human being is explained in brief and then various methods for representing voice in computers are explained, i.e. Parametric methods of voice synthesis, direct waveform coding and synthesis as well as text to speech synthesis. Also their advantages and disadvantages are discussed. The various factors which are used to measure the performance of the particular method are also explained.

In chapter 3, different voice technologies which are coming up and can be used in office automation are discussed.

This is followed by the explanation of an application "A Railway Announcement System", designed to illustrate the use of voice technology (particularly, voice response systems) in office automation, in chapter 4.

Chapter 5 lists the conclusions and recommendations.

CHAPTER 2

VOICE FOR COMPUTERS

2.1 INTRODUCTION

Representing voice in digital form allows the application of digital computer technology to the processing of voice signals.

Many techniques exist for voice representation in computer, ranging in complexity from simple sampling and encoding of waveforms to estimation of parameters from a human model of speech production. When choosing a particular representation method, the various factors need to be considered. These factors are :

1) **Processing complexity** - which refers to the amount of processing required to obtain a digital representation from analog signals and vice versa. It is a measure of cost of implementation of the system in hardware and/or software.

2) **Data rate** - refers to the rate at which digital data has to be transferred from memory to the decoder or synthesizer to reproduce the speech signal (decoder and synthesizer are explained later). A low data rate will result in the low storage requirements and capability of transferring more information over a channel.

3) **Speech Quality** - is a measure of how well the reconstructed signal approximates the original signal.

4) **Flexibility** is a measure of the degree to which the various attributes of the speech (e.g. pitch, speech rate and quality) can be manipulated independently of one another. It is not always possible to control each attribute independently.

In general any technique should have low processing complexity, low data rate, high speech quality and high flexibility. But these four goals are conflicting and compromises are to be made. Now a days since cost of hardware is dropping rapidly and special purpose digital signal processors are available, processing complexity is not prime consideration. Instead, storage requirements, despite the falling price of memory are the most important considerations, when system has to represent a large number of utterances. The choice of one set of parameters over another should be made depending upon the application.

In the following sections the different methods for representing voice in computer are discussed. Since most of the methods use the model of human speech production system, first the human speech production system has been discussed in brief.

2.2 HUMAN VOCAL SYSTEM

The major parts of the vocal system are shown in the fig. 2.1. The vocal tract is a non-uniform acoustic tube about 17 cm in length (for an adult). It is terminated at one end by vocal cords and at other end by lips. The cross sectional area of the tract is determined by placement of the lips, jaw, tongue and velum (Velum is the soft palette which separates vocal tract and nasal tract) In non-nasal sounds the velum seals off the nasal cavity and no sound is radiated from nostrils.

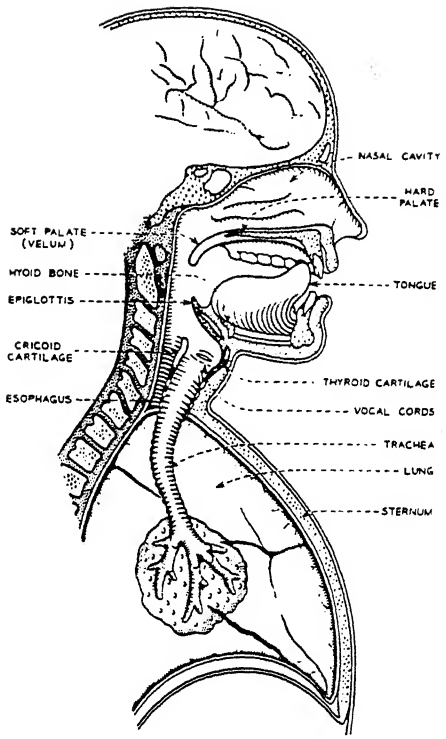


Figure 2.1

Schematic Diagram of Human Vocal System

The sound is generated as a result of an acoustical excitation of the vocal tract. The excitation source is different for different groups of sounds. Voiced sounds (such as vowels /a/,/e/) are produced by vibration action of vocal cords. These sounds are produced by elevating the air pressure in the lungs, forcing a flow through vocal cord orifice (also called glottis) and causing cords to vibrate periodically. Variable area orifice produced by vibrating cords permits the almost periodic pulse train of air to excite acoustic system above vocal cords. Waveform of the glottal volume flow for a given individual vary widely. It depends upon sound pitch and intensity. Fricative sounds (such as /s/,/sh/) are produced by excitation of the vocal tract provided by turbulent air flow passing through constrictions formed at some point in vocal tract usually towards the mouth end. Noise source of sound pressure is thereby created. Plosive sounds result from making a complete closure, again usually towards the front end, building up the pressure behind the closure and abruptly releasing it. All these sources are relatively broad in spectrum. Vocal system acts as a time varying filter to impose its spectral characteristics on the sources.

Hence a simple model of vocal system represents vocal tract as a time varying filter excited by a pulse source (for voiced sounds) or a noise source (for non voiced sounds). The variation within time, of the filter are at a rate sufficiently high to approximate the movements of the vocal tract.

If U_M is the volume flow velocity at the mouth and U_G is the glottal volume velocity, then the Fourier Transform of ratio of mouth to glottal current, (U_M / U_G) is the transmission function of the vocal filter. When the amplitude of this function is plotted against frequency (Frequency spectrum), peaks which occur are called formants and the frequencies at

which formants occur are called formant frequencies. These are given by the equation

$$f_n = (2n - 1) c / 4 l \qquad \text{for} \qquad n = 1, 2, \ldots$$

where l is the length of the vocal tract.

For vocal tract length of 17 cm and sound velocity of 340 m/s these frequencies are 500, 1500, 2500, ...Hz.

This transmission function characteristic is the "filter" that operates on vocal sound source.

Every shape of the vocal tract has a unique set of formant frequencies and the distinctive sounds of a language have perceptually distinctive formant positions. For Example (as shown in fig. 4.2) the vowel /i/ (as in eat) has typically a low first formant frequency and a high second formant frequency. By contrast, the vowel /a/ (as in father) has a high first formant frequency proximate to its low second formant frequency. Overall spectral shapes of two formants are also notably different.

In continuous speech, the formant resonances move around as vocal tract changes shape. But the pitch and intensity parameters vary slowly because of physical limitations on how quickly the vocal tract shape can be changed. Hence they occupy only a small bandwidth.

If these resonances can be determined quickly and preferably automatically, they can be used with data about fundamental voice frequency. and intensity to synthesize signals similar to natural speech.

Speech at acoustic level is not particularly discrete. The articulations of adjacent phonemes interact and that transient movement of

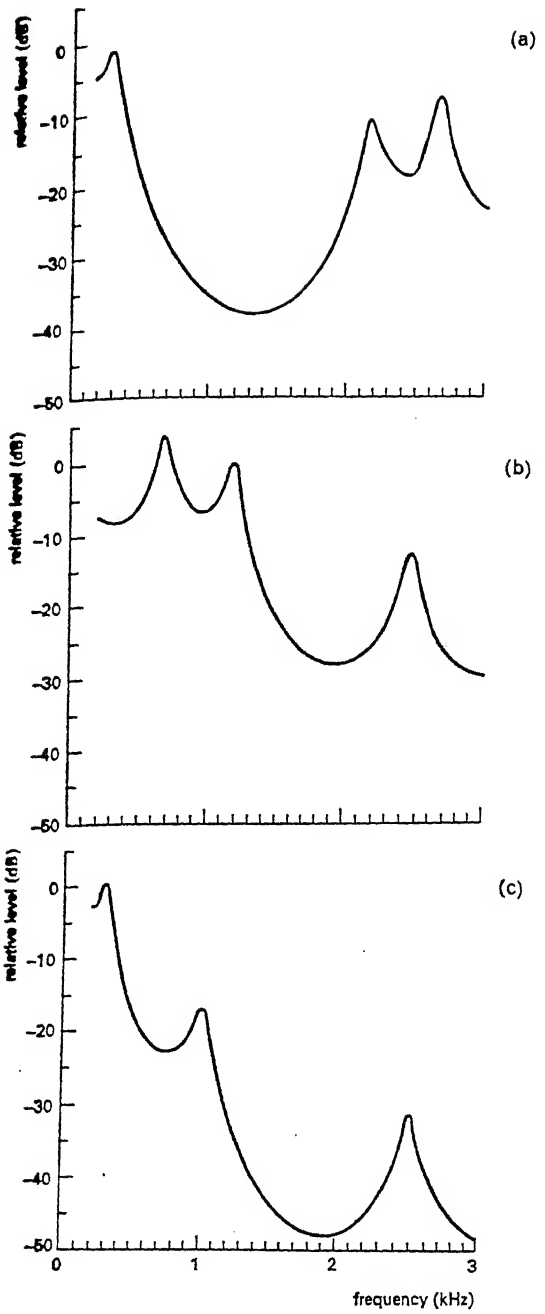


Figure 2.2

Spectrum envelopes approximating the vowels

/i/, /a/, and /u/

the vocal tract for production of any phoneme last much longer than the average duration of the phoneme (total time divided by number of phonemes). That is the articulatory gestures overlap and are superimposed.

2.3 METHODS FOR REPRESENTING VOICE

There are various voice representation techniques for encoding as well as generating the voice from computer. In the next section Parametric encoding techniques are explained and in later sections waveform encoding techniques and Text-To-Speech Synthesis technique are explained.

2.3.1 PARAMETRIC ENCODING

In Parametric encoding method certain parameters that represent the speech signal are extracted and stored. These parameters are determined by the process of speech production. As explained in last section, a simple model of the vocal system represents the vocal tract as a discrete time varying filter excited by a pulse source (for voiced sound) and or by noise source (for non voiced sound). The variation within time of the vocal tract shape can be modeled by varying the parameters of the filter at a rate sufficiently high to approximate the movements of the vocal tract.

The different techniques which are based on this speech model are Formant Synthesis, Linear Predictive Coding (LPC), and Channel Coding. In the following section, Formant Synthesis is explained in detail. In the explanation more stress has been given on the qualitative aspects rather than on mathematical aspects.

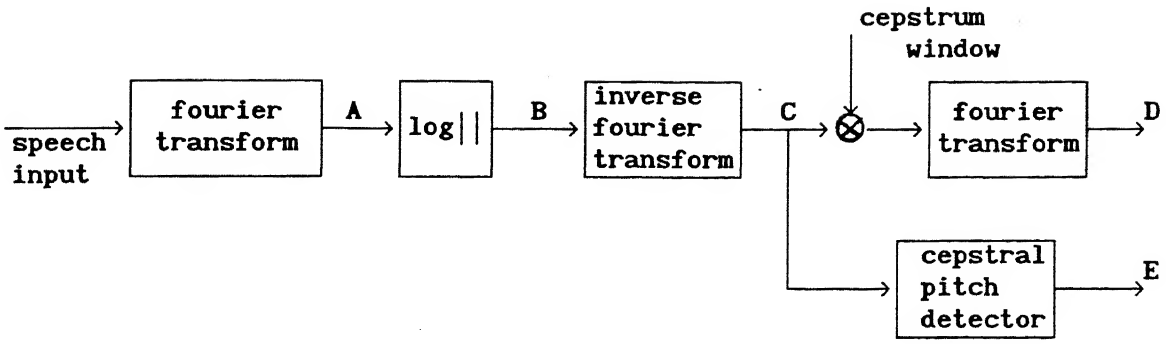
2.3.1.1 SPEECH SYNTHESIS FROM FORMANT DATA

FORMANT ANALYSIS

This method is based on the spectral analysis of a speech signal. Formants, as explained in last section are the resonance peaks of the spectral sections of speech, which are determined by the shape and dimensions of the vocal tract. A sound can be represented by its formant positions, amplitude and pitch period of the excitation. Because these speech parameters vary slowly in time, the properties of the waveform usually remain roughly the same over a short duration (usually 10 - 30 ms). Thus by analyzing short segments of a speech signal, its characteristics can be determined (Short time analysis). The time variation of the waveform during a single period is determined primarily by the vocal tract response, whereas the fundamental period (pitch period) reflects the vibration of the vocal cords. The Fourier Transform of a short segment of the speech waveform (frequency spectrum of a signal) will reflect features of the excitation and formant frequencies for that segment.

Fig. 2.3 [16] gives the basic operation that isolates the features of the excitation source and formant frequencies.

According to the model of the human vocal system, the speech signals are generated by superimposing the vocal tract function and the excitation function, which represent the spectra of the vocal tract and excitation source, respectively. Therefore the first Fourier Transform will give the frequency spectrum of the input signal, which is actually a product of the frequency spectra of the vocal tract and the excitation source. Taking the logarithm of the magnitude of the the Fourier Transform turns the product into a sum of logarithms. The Inverse Fourier Transform gives the time



A - product of frequency spectrum of excitation source and vocal tract function
B - sum of frequency spectrum of excitation source and vocal tract function
C - cepstrum of input signal
D - frequency spectrum of vocal tract function
E - pitch period of the excitation source

Figure 2.3

Short Time Analysis of Speech

domain signal called as *cepstrum* of the input signal which is an additive combination of the cepstra of the excitation and the vocal tract functions. The two cepstra do not overlap in time since the excitation function is periodic and varies at a rate much higher than that of the vocal tract function. To obtain the vocal tract function, the cepstrum is multiplied by a window called the cepstrum window, which truncates the cepstrum values above 4 ms to zero [7]. Taking the Fourier Transform of the zero to 4 ms portion yields the frequency spectrum of the vocal tract function. The peaks in the frequency spectrum (i.e. formants) then can be easily detected by examining the amplitude of each frequency component. The pitch component is mainly represented by sharp peaks in the cepstrum at multiples of the pitch period. Thus pitch period can be determined by searching the cepstrum for the first peak after the 4 ms portion.

The average amplitude of the samples in the segment is also calculated for use in the synthesis process. These operations are performed repeatedly on consecutive segments of speech, yielding a set of parameters to be used in the synthesis process.

After the parameters, i.e. formants, pitch period and amplitude are obtained and stored in memory, speech can be reconstructed by supplying the parameters to a synthesizer that approximates the model of human speech production. These parameters can also be transmitted over the channel and at the receiving end can be used to generate the speech again.

FORMANT SYNTHESIS

Once the excitation and transmission parameters are obtained, they are used to synthesize a waveform that approximates the original speech signal. A number of systems have been devised for formant synthesis, like terminal-analog speech synthesizer, OVE synthesizer, digital formant synthesizer. The digital synthesizer has been shown in fig. 2.4 [7]. Its

excitation source is a impulse train with spacing equal to fundamental pitch period. The amplitude of the pulse signal A_V, also estimated from natural speech, controls the intensity of the pulse excitation applied to cascade of variable digital resonators. The output of this system excites the system designed to approximate the spectral shaping due to radiation and source properties. The lower branch produces the unvoiced speech. A random noise generator, whose intensity is given by A_N, excites digital filter. The output of this system passes to fixed spectral compensation filter to provide the unvoiced speech output. Control parameters are supplied to the synthesizer at least at their Nyquist rate. (Nyquist rate is explained in next section.)

To synthesize the continuous speech, timing, pitch, and formant information must be generated. Timing information is derived in several ways. The techniques are

1. External specification of the duration of each word is chosen according to some external criterion. (e.g. it can be measured from naturally spoken version of the message.)
2. Calculation of word duration by rules based on models of English language.
3. Specification of word duration from tables of stored data. For limited context messages, such as sequences of digits, such specification of word duration often is acceptable.

To determine the pitch contour (i.e., the pitch period as a function of time) is the next step in synthesizing the message. Pitch information can be added by supplying a pitch contour determined from the originally spoken message. Data of this would normally be used when word durations have been obtained in a similar manner and supplied externally. Another

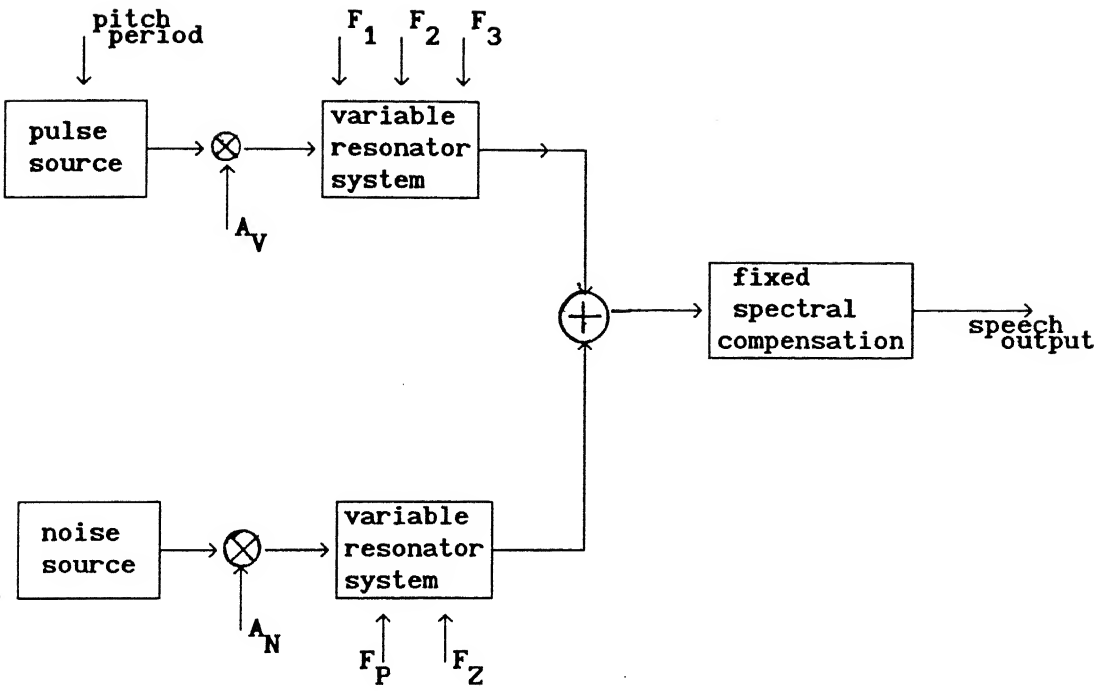


Figure 2.4

Architecture of Speech Synthesizer using Formant Synthesis

alternative is calculating a pitch contour by rule [7]. Once the timing pattern is obtained, formant data must be altered to match the timing. The formant contours in successive words must also be merged together to form the smooth, continuous transitions. The final step in producing the message is to synthesize the speech using the chosen prosodic features and segmental features generated by the preceding rules. A hardware digital speech synthesizer performs this task in real time at a rate of 10 KHz.

Fig 2.5 [7] gives the block diagram of concatenation program.

2.3.2 DIRECT WAVEFORM ENCODING

This has been the earliest approach taken for representing voice in computers. Nyquist's sampling theorem states that any band limited signal can be exactly reconstructed from samples taken periodically in time if the sampling rate is twice the highest frequency (called Nyquist frequency)of the signal. (This rate is called the Nyquist rate.) Waveform encoding techniques are based on this important theorem. There are large number of different wavefrm encoding techniques. Some of those are Pulse Code Modulation (PCM), Differential Pulse Code Modulation (DPCM), and Adaptive Differential Pulse Code Modulation (ADPCM). DPCM technique has evolved from PCM and ADPCM technique from DPCM. In the following sections these methods are explained.

Before explaining the above mentioned techniques, it is necessary to understand some fundamental concepts of digital signal processing which will help to understand those techniques.

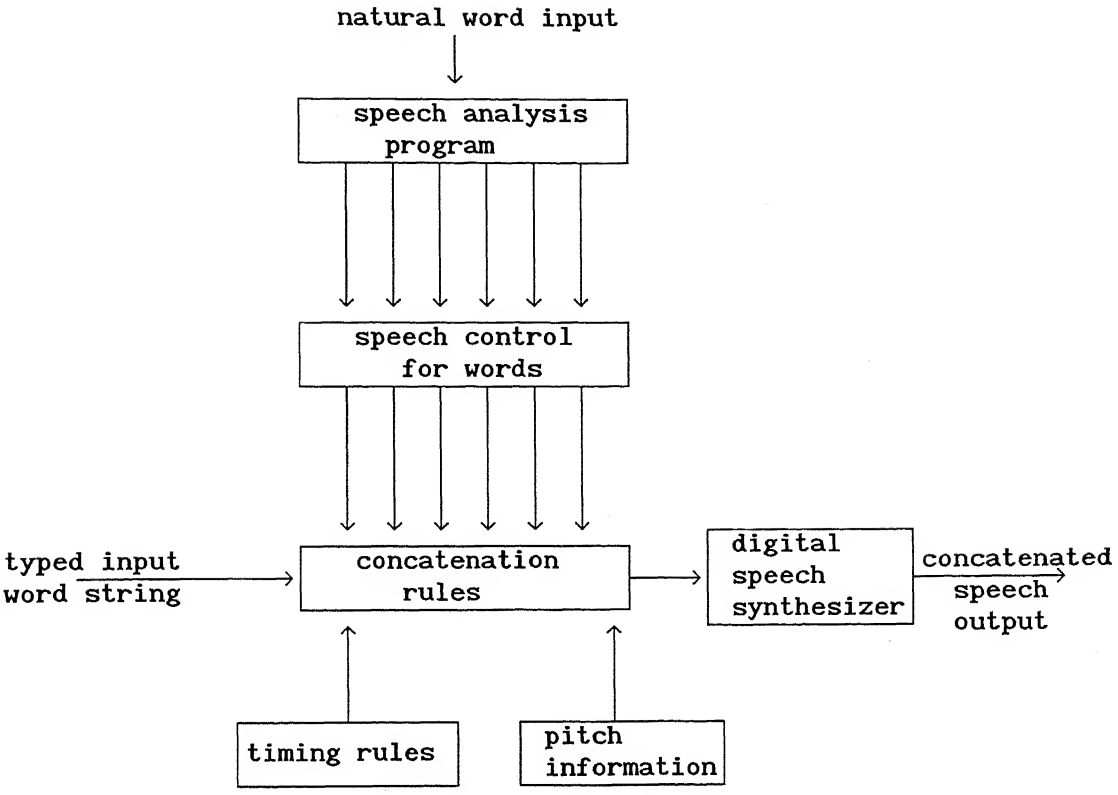


Figure 2.5

Block Diagram of Concatenation Program

2.3.2.1 FUNDAMENTALS OF DIGITAL SIGNAL PROCESSING

Generally human ear can hear sounds in the range of 20 Hz to 20,000 Hz., while human speech is in the range of 300 to 3000 Hz. In the fig.2.6 an audio waveform is shown. The waveform is a combination of various frequencies at different amplitudes and phases. (According to Fourier Theorem, any periodic signal can be described as the sum of single frequency sine waves of various amplitudes.). This figure represents a 5 ms portion of waveform. The first step is sampling, which consists of taking measurements of the input signal at regular intervals of time, next step is quantization i.e., converting the samples to appropriate scale, and storing them. Fig 2.7 shows the same waveform sampled at 8 KHz sampling rate i.e., 8000 times per second a measurement is taken of the input signal. In most computer systems, the sampling and quantization is done with an Analog to Digital Converter (ADC). An eight bit ADC might be having range of output from -128 to +127 for input voltage range of -250 mV to 250 mV will output an 8 bit value of 127 if the input voltage is 250 mV and -128 if the input is -250 mV. Once the sound is digitized, it can be stored in the computer memory. To play back the sound, the Digital to analog Converter (DAC) is required. DAC takes the digital value and converts it to a corresponding analog signal. While the exact voltages produced at the output of the DAC do not need to be identical to those at the input, they do need to be proportional to one another so that one waveform corresponds to other. Also the samples need to be output at exactly the same rate that they were read in. Fig 2.8 shows the output of the DAC for the samples of the fig. 2.6. The sharp jumps that occur when moving from one sample to another represent the high frequency components in the signal, these can be removed by means of passing the signal through a low-pass filter.

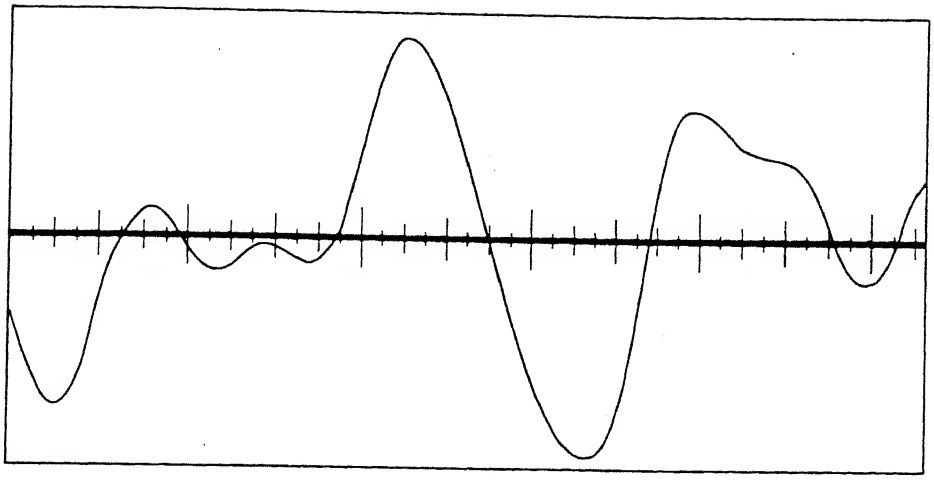


Figure 2.6 An Audio Waveform

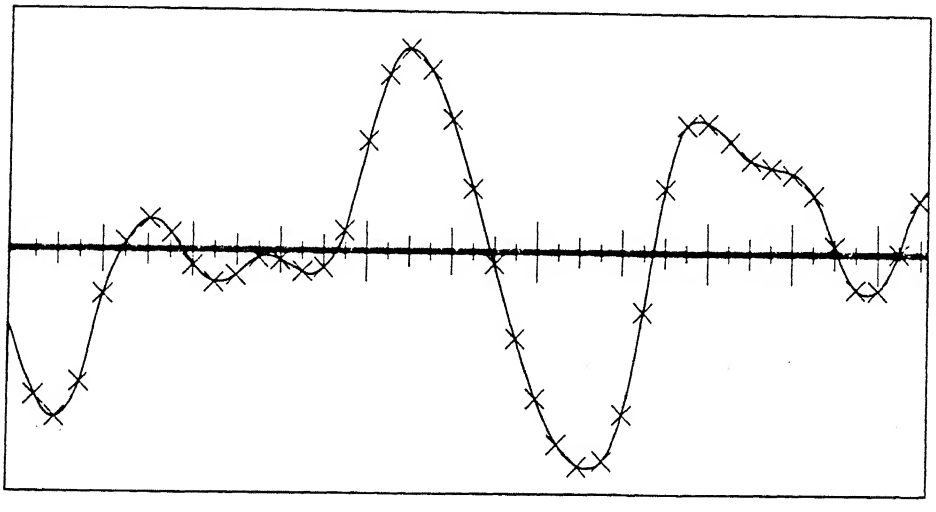


Figure 2.7 An Audio Waveform Sampled at 8 KHz

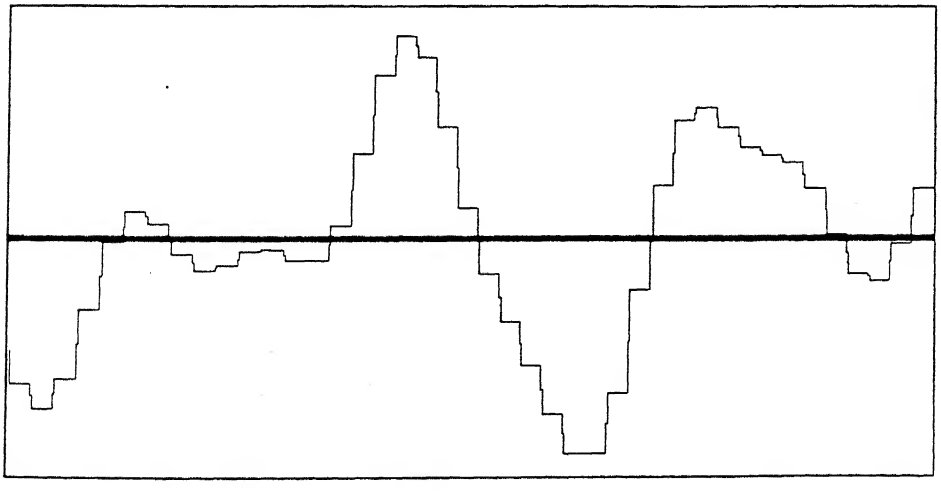


Figure 2.8 DAC Output

When an audio waveform is sampled, two important variables affect the quality of reproduction; the sample resolution and the sampling rate. Sample Resolution is simply a measure of how accurately the digital sample can measure the voltage it is representing. In the quantization step, while converting a input range of -500 mV to 500 mV, an eight bit ADC can resolve the input signal to about 4 mV. (This is also called the step size.) Hence an input signal of 2 mV will either get rounded up to 4mV or down to 0 mV. Clearly, qunatization loses the information and introduces quantization noise into the signal. The greater the number of bits representing the input, obviously less will be quantization error. Sampling Rate plays a different role in determining the quality of digital sound reproduction. According to Nyquist theorem (as mentioned earlier) a 20 KHz bandwidth audio signal must be sampled at 40 KHz for it to be properly reproduced.

If a signal is sampled below the Nyquist rate, the sample points do not contain enough information to reconstruct the original signal. Frequencies in the input signal above Nyquist frequency generate undesirable frequencies in the digital signal. (This is called Aliasing.) For example, if there is 22 KHz signal component in the input signal before it is sampled at 40 KHz, the digital signal will contain an 18 KHz signal but not 22 KHz signal. To remove the undesirable frequencies, the frequency components above the Nyquist frequency must be filtered out using low pass filter before it is converted to the digital signal.

Human ear hears sounds up to 20 kHz, which implies that we need to sample audio at 40 KHz or better to achieve good reproduction. In fact, the sampling rate used for digital reproduction of music via compact disk or digital audio tape is 44 KHz, using 16 bit samples. The quality of sound achieved at this sampling rate is generally acknowledged to be

superior. But to achieve the digital signal of human speech signal (frequency up to 3 KHz), the only sampling rate of 8 KHz is required. In fact every digital phone system in the world uses an 8 KHz sampling rate to record human speech.

These digital signals can be stored on hard disks for editing and playback. We can estimate the storage requirements required to handle stereo quality audio. For sampling rate 44 KHz, using 16 bit samples, to store an hour of music takes over 600 MB of storage. Hence the need of data compression arises. Signals can be compressed or coded to reduce their large storage requirements. There are two basic types of compression schemes, lossless and lossy compression. Lossless compression is used when a reconstructed signal must be same as the original signal. Lossless compression algorithms can usually compress digital data up to one-half. or one-quarter of its original signal. Some lossless compression algorithms are Huffman and Lempel-Ziv [28].

Lossy compression algorithms are used when the reconstructed signal does not have to be identical to the original signal. This is the case with audio, video and graphics compression. These data types often have more information than that can be perceived by human receiver. Thus compression algorithms can afford to lose information . Lossy compression can frequently compress digital data to as much as one tenth to one hundredth or more of its original size. Lossy algorithms include PCM, DPCM, ADPCM. Since ADPCM is the algorithm which is most oftenly used in audio compression, this is explained in the following section.

2.3.2.2 ADAPTIVE PULSE CODE MODULATION (ADPCM)

Pulse Code modulation is the simplest way of speech coding. The speech signal is first filtered to a bandwidth of about 4 KHz to limit its highest frequency and then sampled at the Nyquist rate of 8 KHz. Stereo quality sound is generally first filtered to a bandwidth of 20 KHz and sampled at 44 KHz rate. The amplitude range of interest is divided into a finite number of levels each of which is represented by 8 bit or 16 bit number. The difference between consecutive levels is called step size. Each sample is rounded to the nearest level by a quantizer and is assigned the corresponding PCM code.

While, in DPCM the difference between magnitude of the successive samples is encoded, rather than encoding the magnitude of the samples itself. Since the amplitude of the speech signal varies slowly, the difference in magnitude of successive samples is small. Hence encoding needs a small number of bits. If the sampling rate is increased, the difference between successive samples will be smaller and number of bits required to encode one sample will also less.

When the sampling rate is high, the correlation between the successive samples is also very high. By using this property of a speech signal, it is possible to predict a sample based on a linear combination of previous samples. In ADPCM technique, the use of the same fact is made. In ADPCM, like DPCM the difference signal is first quantized; the quantizer obtains information about the input signal by keeping the tract of previous samples, which are used to predict the next sample. If it is predicted that the input signal is varying slowly, a small step size is used; if not, a large one is used. In this way a quantizer adopts itself to the difference signal.

Fig. 2.9 [6] is a block diagram of ADPCM coder and decoder. It follows the conventional DPCM with the predictor in the feedback loop. The box labeled LOGIC provides the adaptation of quantizer step-size on the basis of most recent output.

ADVANTAGES OF ADPCM METHOD :

1) A low data rate is achieved; at a rate of 24,000 bits per second, ADPCM speech is found to be perceptually comparable to speech quality of the public telephone systems [16].

2) An interesting feature of ADPCM is its ability to determine the beginning and end of an utterance, which is useful for speech editing.

3) Hardware implementation is relatively straightforward [6].

2.3.3 TEXT-TO-SPEECH SYNTHESIS

Synthesis from text offers virtually unlimited message capacity.A block diagram of text to speech conversion program is shown in the fig. 2.10 [7]. Blocks A through D convert text to discrete symbols representing phonemes with detailed data for pitch and duration. Blocks E to I convert the discrete phonemic symbols to continuous sound. Blocks A through C generates the minimum information for pause and stress assignment. The block D generates pitch and timing assignments for each phoneme and produces discrete symbols that are used directly as the input to the articulatory model.

The dictionary provides a phonemic transcription of each word with lexical stress marks, and the possible usages of word such as noun, adjective or verb. The role of syntax analyzer (block A) is to assign for each sentence a probability of break, at every grammatical boundary, to select alternative pronunciation of certain words according to their usage

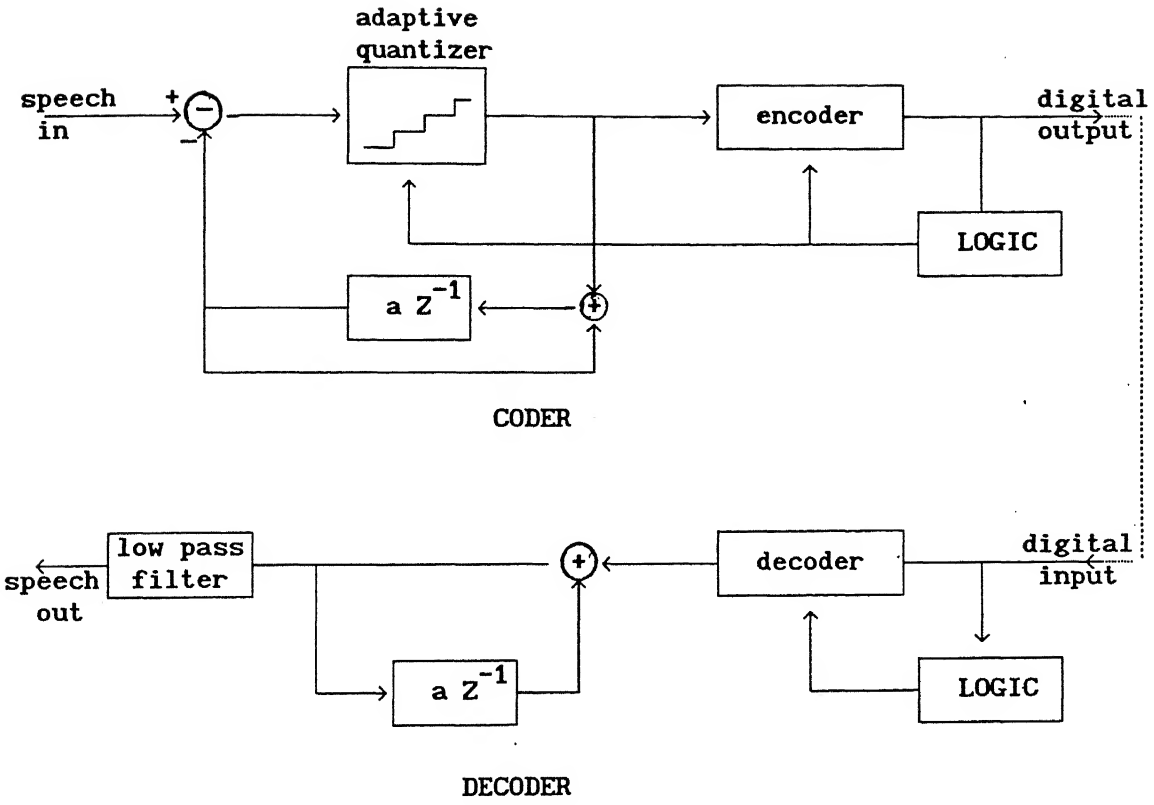


Figure 2.9

Block Diagram Of ADPCM Coder

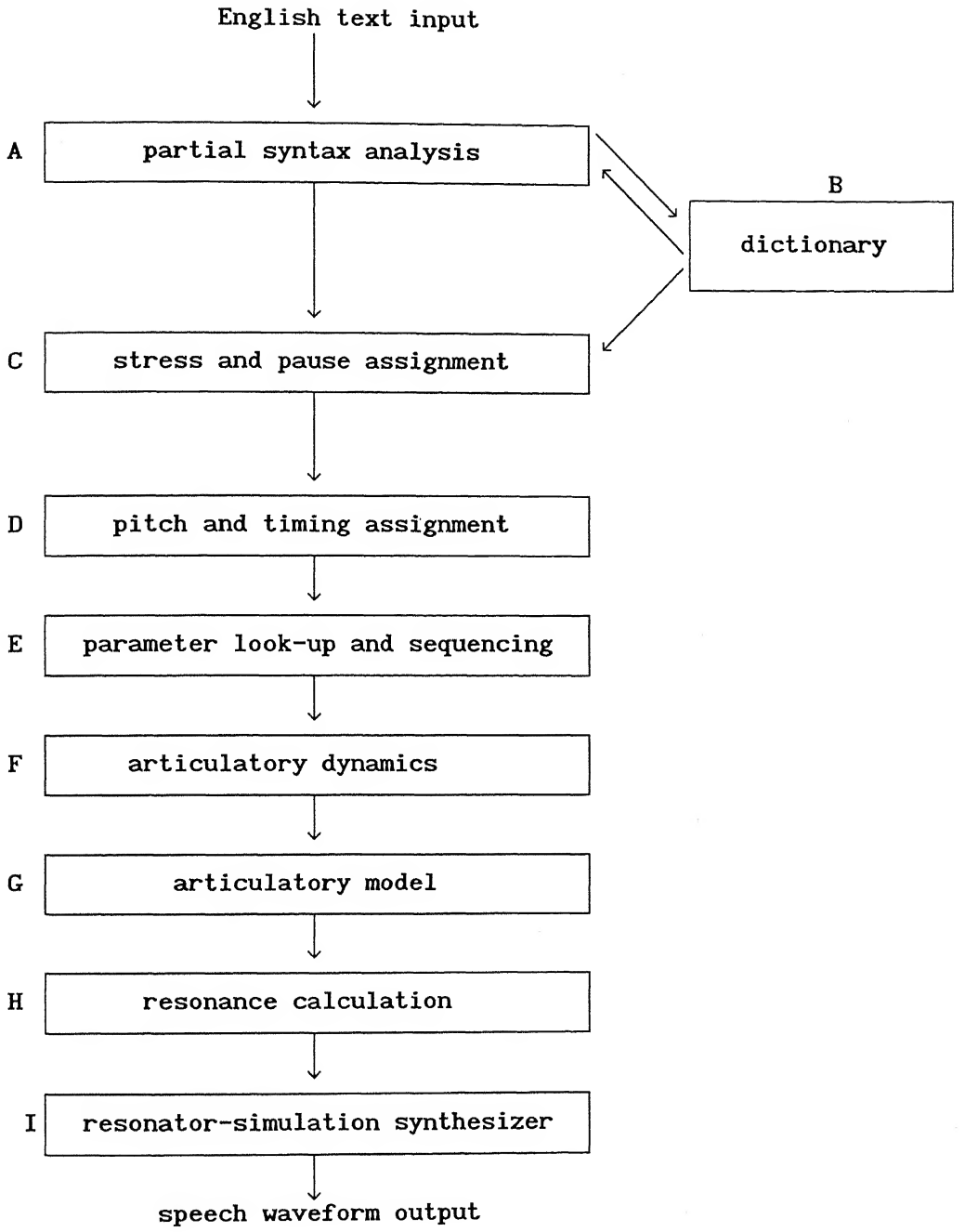


Figure 2.10

Block Diagram of Synthesis of Speech from English Text

and to provide content-function distinction. (Content words convey meaning in the sentence. They relate to things, actions or attributes. Function words serve mainly to establish grammatical relationships. Function words include articles, prepositions and conjunctions and pronouns) Using the information on word class contained in the dictionary, the analyzer groups the words into phrases. A weaker break is assigned between all words. Words in the same phrase have zero probability of break. Break probability is higher between object and predicate than between verb and object. Punctuation marks require the highest probabilities of the break.

Using the information on probabilities of a break obtained in the syntax analyzer, block C, decides what kind of break and pitch contour is to be assigned at the end of grammatical unit. The threshold for putting an actual pause in the synthetic speech depends upon length of the sentence and the speech rate. Full stops indicate the longest pause and have a falling pitch contour at the end, commas indicate a pause and a rising pitch contour, implying continuation. Stress levels are assigned words according to content-function distinction. The stress levels are assigned to words used in computing pitch and duration information.

From all these information, timing-control marks and pitch marks are assigned to each phonemes. Two major rules are used to determine the pitch and timing control : (1) a word boundary rule, which determines consonant durations, and (2) a stress and termination rule, which determines pitch marks and vowel durations. (For these rules refer [7]) The pitch and timing assignments complete the information derived by text-conversion program. The resulting data are then supplied to the articulatory model and converted to connected speech. (Refer [7] articulatory model.)

CHAPTER 3

VOICE IN OFFICE AUTOMATION

3.1 INTRODUCTION

As explained in chapter 1, Office Automation is use of information technology for Gathering, Storage, Retrieval, Processing and Communication of information for improving the effectiveness of office. Voice can play an important role in Office Automation. The different applications of the voice are explained in Chapter 5. This chapter is mainly concerned with the different technologies of voice in Computers. These different technologies are

1) Voice Response
2) Voice Recognition
3) Voice Annotation
4) Voice Messaging.

These technologies are explained in the following sections. Voice Response has been explained in great detail, since it is the subject of the thesis, while other technologies are explained relatively in brief. The benefits of voice in office systems are obvious. In the office, speech is the most oftenly used and natural mode of communication. Managers also feel comfort in speaking and dictating rather than typing for himself. Voice can play a role to alert the person, when he needs to be alerted, if the electronic calendars and reminders are used.

3.2 VOICE RESPONSE

Voice Response is the process of generating human like voice from machine.

Voice response systems are designed to respond to a request for information or to generate a request using spoken messages. Voice Response systems thus allow one-way communication from machine to man.

Voice response systems in general can be classified into the two types :

1) Systems with Limited Vocabulary and

2) Systems with Unlimited Vocabulary.

In the following sections several types of voice response systems, their merits and demerits are explained. Before explaining the voice response systems, some factors need to be considered to measure the performance of the system. Hence these factors are explained first.

3.2.1 FACTORS TO MEASURE PERFORMANCE OF VOICE RESPONSE SYSTEMS

1) Size of Vocabulary : Size of vocabulary refers to the number of words, phrases, or sentences that the system can speak. In some applications such as speaking clock or speaking calculator a fixed number of phrases can be stored in the system. But in some applications such as a verbal response of an information retrieval system, the response must be able to produce a large, if not infinite number of utterances.

2) Speech Quality : Speech Quality is mainly a subjective measure, ranging from merely intelligible, discrete word pronunciation to natural sounding speech that has realistic features. As we have seen in last

chapter various voice synthesis techniques have been designed to produce the speech.

3) Cost of System : The cost of a system is usually measured in terms of the storage requirements. Obviously, systems with limited vocabulary and low speech quality cost the least, whereas systems with unlimited vocabulary and high speech quality cost the most. In practice, systems with unlimited vocabulary have to sacrifice speech quality in order to make the system economically feasible.

4) Flexibility : The flexibility of a voice response system refers to how easily existing utterances can be modified and new utterances added to the system.

3.2.2 TYPES OF VOICE RESPONSE SYSTEMS

3.2.2.1 VOICE RESPONSE SYSTEM WITH LIMITED VOCABULARY

In such a system, the vocabulary is limited to the number of words, phrases or sentences, encoded in the system beforehand. The encoding and synthesis for a system with limited vocabulary is shown in figure 3.1.

In the encoding phase, an utterance is spoken into the system and is subsequently converted into digital form by either direct waveform coding or parametric coding. (Waveform coding and parametric coding has been discussed in the chapter 2.)

The speech parameters extracted from parameter coding or digital samples encoded by waveform encoding are then stored into memory for later retrieval.

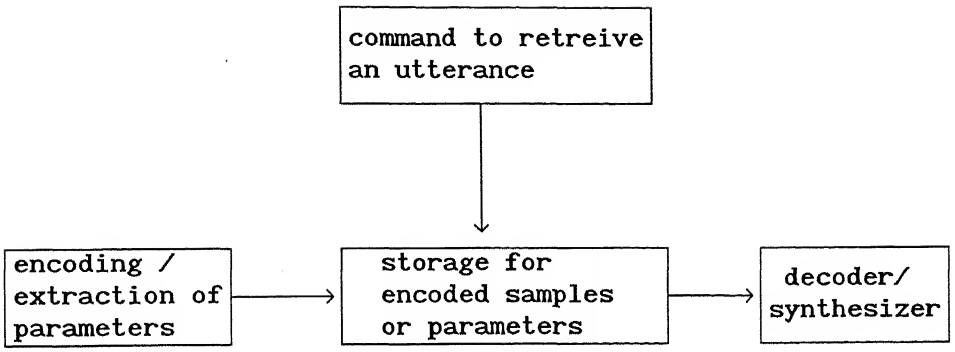


Figure 3.1

Limited Vocabulary Voice Response System

In the decoding phase, command to retrieve an utterance is given to the system, which determines the storage location where the utterance is stored. Then the parameters or encoded samples are retrieved and fed to the synthesizer or decoder.

In fact this method is merely a digital version of conventional analog recording method.

Since in this system, vocabulary is limited, the limitations of such system are obvious. Moreover, in order to modify existing utterances or add new utterances to the system, the encoding process must be invoked. Also the new utterances must be spoken by the same person in order to keep the quality compatible with the existing utterances.

The important benefit of such a system is the high speech quality that can be achieved as a result of direct encoding of a utterance and even the identity of the speaker can be retained.

Memory requirement of such a system depends upon encoding method and varies from 1.2 Kbits/second for Linear Predictive Coding (LPC) to 70 Kbits/second for direct Pulse Code Modulation (PCM). (Different methods have been explained in chapter 3.)

3.2.2.2 VOICE RESPONSE SYSTEM WITH UNLIMITED VOCABULARY

One way to achieve a system capable of producing an infinite number of utterances is to store all possible utterances in the system. Even if this approach were possible, it is not feasible to produce, since the high cost of storage and high processing time would not allow it.

Voice response systems with unlimited vocabulary can be produced in either of the following two ways :

1) Phonetic Synthesis

2) Text to Speech Synthesis

PHONETIC SYNTHESIS

Phonetic synthesis is based on the principle that any word of a language can be represented by a small number of linguistic units such as phonemes, allophones, syllables. The representation of these smaller units require considerably less memory. A phoneme is the smallest segment of sound that distinguishes words of different meanings in a language. For instance /p/ and /b/ are the phonemes of English. If one is replaced by the other, a different meaning will result, as in the words "pin" and "bin". Some phonemes have different sounds, depending on the context in which they occur. For instance, the sound of /p/ in "pin" and sound of /p/ in "spin" are variants of the same phoneme /p/. These variants of phonemes are called allophones. Allophones are important for capturing high quality speech. A morpheme is the minimal unit of sound carrying meaning. For example, "dogs" consists of a root morph "dog" and an inflectional suffix "s" which denotes a plural.

In this system, phonemes are first encoded and stored in memory. To create an utterance, the utterance is first transcribed into its phoneme representation. In some systems prosodic information is also embedded into the transcription (e.g. special characters are used to mark the stress patterns.) in order to produce more natural speech.

Languages contain relatively few phonemes; the storage requirements are therefore small. Also a word can be represented by a small number of phonemes, hence the data rate is also low. However phonetic synthesis based on phonemes and allophones requires complicated rules to compute the transition between two adjacent phonemes or allophones, required to achieve the intelligible speech.

TEXT-TO-SPEECH SYNTHESIS

In phonetic synthesis the text to be spoken is first converted into its phonetic representation. In Text-to-Speech synthesis system, first the text is directly translated to speech. This is usually done in two steps. First the text is transcribed into its phonetic representation using a set of linguistic rules. Second, phonetic transcription is used to drive a phonetic synthesizer in which transcription is translated to speech parameters. Although Text-to-Speech synthesis is the most complex form of speech synthesis, it is very useful since it enables any text stored in a computer to be spoken.

A Text-to-Speech translation algorithm can be composed only of a set of letter-to-sound rules, or it can include reference dictionary. Letter-to-Sound rules attempt to guess the pronunciation of letters according to pronunciation rules. Look-up dictionaries are usually used for handling exceptions. Depending on the system design, the dictionary may also carry other information such as the stress pattern of the word. If a word is found in the dictionary, no translation by rule is required. (In chapter 2, Text-to-Speech synthesis has been explained in detail.)

In the Text-to-Speech synthesis, text is required to be stored in the computer. But many information systems do not store the complete text. For instance, an employee file is usually stored as tuples with different field such as employee name, salary; instead of storing complete sentences to be output. In this case, when a Text-to-Speech system is used, a sentence is generated from the data stored. The text then is passed to a text-to-speech system which converts it into speech. (This is called speech synthesis from concept) The objective of such system is to provide the speech output for an information system to avoid the redundancy.

3.3 VOICE RECOGNITION

Voice recognition involves inputting data to a computer using a human voice rather than using a keyboard or other electromechanical means. The aim is to enable computers to listen to (and understand) human speech. The ultimate objective is to develop a system capable of understanding people as they speak normally. This is known as speaker independent continuous speech recognition. Currently the state of the art falls some way short of this.

DIFFERENT CATEGORIES OF VOICE RECOGNITION SYSTEMS

The voice recognition systems can be categorized as

1) Speaker Dependent systems and

2) Speaker Independent systems.

The speaker dependent recognition system requires 'training' by the individual user whose voice will be used to operate it. This training is carried out by the user entering a set of vocabulary words. In training, the voice system digitizes and creates a template of each word and stores the template in system memory. In operation the voice system digitizes the a spoken word and compares it to the stored templates. If they match, the word is recognized. These systems can not be used by other than one recognizer speaker at a time.

Defining the vocabulary for the speaker-dependent system requires several steps. Firstly the user must determine the words he wants the system to recognize and then type these words into the system. These typed words are used as a prompt when user is teaching the system. The next stage is to decide upon the action to be taken by the system upon recognizing the word. Typically, one vocal command will control a number

of keystrokes. Thus it is possible for one voice command to activate a whole series of commands that would normally be entered manually. The third stage is to teach the system to recognize the user's voice. The system prompts the user by pointing to each word in the vocabulary in turn, the user responding by speaking each word; since it is almost impossible to say any word exactly the same way twice, the process is then repeated up to five times to average the variations in the way a word is pronounced. Most systems incorporate a method of subsequently testing the degree to which the system recognizes the user's voice.

SPEAKER INDEPENDENT VOICE RECOGNITION SYSTEMS

These systems are not limited to a single speaker. these systems have the capacity to identify a wide range of speakers without the necessity for developing individualized vocabularies. The speaker training is replaced by a set of representative templates created from a statistical sampling of a large population base, for differences in the accent, tone and speed. The major drawbacks of such systems are that

1) the vocabularies are limited;

2) the vocabularies are preselected by the vendor.

speech recognition technology is still well short of speaker independent continuous systems. Currently available voice recognition systems are categorized as ;

1) Speaker Dependent, Discrete Word Systems : These are the most prevalent systems because they have the highest accuracy and the lowest cost, some capable of speaker dependent, continuous word operation.

2) Speaker Independent, Discrete Word Systems : These are mainly used in the telephone applications where many different speakers will be encountered,

3) Speaker Dependent, Continuous Word Systems : Such systems are new beginning to emerge onto the market.

VOICE RECOGNITION

Most of the voice recognition systems work by first establishing the pattern of the word. The spoken word is then matched to this voice patterns already stored in the memory. Voice recognition is a complex process, which consists of following steps ;

When the user speaks, using microphone the sound wave is converted into the analog signal. This signal is then digitized at the sampling rate of 8 KHz using 8 bit or 16 bit samples.

In discrete word recognition, the end point of the word is determined by the period of silence which follows it. In case of continuous speech recognition this is much more difficult since people tend to speak words together. Therefore, techniques incorporating knowledge of language syntax have been developed to try and overcome such difficulties.

Once the sound wave is converted to a digital form, the recognition system then uses a variety of mathematical techniques to break it down into a pattern of intensity, thus creating a word template. (Full mathematical details are beyond the scope of this thesis.)

The system then tries to match the template of the word with the template of those words contained within the predefined vocabulary. After finding the best match, the system decides if the input word is close enough to the best match to justify a positive recognition. If the system

decides that a recognition justifies, if will notify the user either via the terminal's display or using a voice synthesis facility.

There are many difficulties associated with recognition process, like background music, which is gradually being overcome. Another difficulty is the fact that current systems merely recognize words as opposed to understanding them within the context. For example people may have difficulty in differentiating a 'our' from a 'hour', but are able to overcome such ambiguities through the context in which the word is spoken. Voice recognition systems can not understand the words, and resolve ambiguities.

Another problem concerns the voice of the speaker. The human voice is the most inconsistent transmitter. Emotion completely changes the way you speak. If you are tired, The same words come out slower and lower; if you are excited or angry, you speak faster. Further difficulties exists for the recognizer if the speaker has a bad cold. Speaker independent system are likely to have problems where speakers have heavy regional accents.

3.4 VOICE ANNOTATION

Annotation of paper documents is a basic office function. Any comments made are underwritten, usually against the piece of the text they refer to, and also carries the signature of the person making the comment.

However the advent of the electronic document, particularly electronic mail, has made annotation a much more difficult operation. Using a keyboard to annotate is problematical because of two reasons :

1) The format of the document may not allow space to annotate at the appropriate point and

2) An electronic annotation on an electronic document does not stand out as being such. The handwritten comment on the paper document stands out very clearly as having being added at a later day. Some word processing systems unable the annotation to be printed out in a different color which solves the problem for the hard copy version, but does little for the electronic version shown on the screen.

This problem is now being addressed by voice annotation. This facility allows the user to attach a vocal comment to an electronic document. Thus the user of an office workstation is able to 'tag' voice the comments onto letters, memos or reports. The voice comments become a part of the document and may be stored or transmitted within compound document. When the document is retrieved the voice comment can be listened to in order explain the textual portion of the message.

A special character is normally placed in the electronic document to show the position of a voice annotation. This special character may be placed within the body of the text or opposite the line to which it refers. The user can access the voice component by placing the screen cursor on that character and requesting the system to play back the message. The different applications where the voice annotation can be used are explained the chapter 5.

3.5 VOICE MESSAGING SYSTEMS (VMS)

Voice messaging systems are also called as "store and forward voice message systems". Voice messaging is non-simultaneous verbal communication. It is similar to electronic mail (text), except that audio signals are processed rather than text. A sender's voice signal from a telephone (or through a microphone) is digitized, compressed and then stored for subsequent processing. The message can be delivered at a

specific time in the future, or retrieved from a voice mailbox by the recipient. The message is reconverted to its analog form when it is retrieved, and fed to the speaker so that it sounds like the original sender's voice. The basic operation of a typical Voice Messaging System can be explained as follows.

To send a message the speaker will log into the system by writing a login id and password, and will be connected to the system. He will then give command to record the message and starts speaking. To stop recording the message another command will required to be given. Voice message, thus generated is stored as a sound file. With some commands the message can be edited, i.e. some portion of the original message can be deleted, some portion can be included, or different messages can be merged. Once the sender is happy, he can give command and receiver's login id to send the message to the recipient's voice mailbox. In case of VMS using phone, to send the message, the user will dial the voice message facility. This facility may be part of his local PABX. Once connected to the system, user identifies himself himself by dialing his identity number. Sender also dials the recipient's mailbox address. Then a single digit command will be given to start, stop recording, and thus generating the message. Once message is created, another single digit commands will allow to edit the message. When the sender is ready, by giving a command (i.e. dialing a one digit number). will send the message to the recipient's mailbox.

To receive message from the mailbox, a recipient needs to identify himself as above. By using appropriate commands the user can listen particular message, he can skip some portion of the message, he can store the message for the future use or delete it. Having listened to the

message the user can then send reply to the to the originator (without having his address), or can redirect a message to others.

The motivating factor behind the emergence of the VMS is the increasing need in today's business to communicate information as quickly and as accurately as possible. However, a major problem with communications between business people is that they are not always readily available; for instance they may be travelling, in meetings, away from the desk, at a different location, in a different time zone or simply not wanting to be interrupted.

As mentioned above the two major methods by which two facilities are provided are

1) PABX based and

2) Workstation based.

PABX based VMS systems fall into two categories depending upon the method used to connect the system to the telephone equipment. The categories are stand-alone and integrated.

In a stand alone VMS, in order to use a stand alone system the user must call system number. Thus if the user called the recipient over conventional telephone line using the ordinary telephone number and did not receive a reply, he would have to re-dial the VMS system number. This is because the PABX is not capable of passing addressing information to the VMS in order to allow the call to be automatically re-routed to the mailbox. The advantage of the stand-alone system is that the system will normally be PABX independent.

In integrated VMS, a data path is available between the PABX and the VMS. This in turn allows two-way communication so that, addresses can be passed from the telephone system to the VMS, or the PABX can be notified of 'waiting messages'.

The workstation based VMS permit voice messages (or mixed messages consisting of text, graphics and text) to be communicated to other users on the same system. Currently these systems cover relatively small user communities.

Features and Benefits derived from the VMS are as follows :

1) VMS does not require the recipient to be present in order to receive the message. It thus eliminates the wasteful necessity of playing telephone tag when trying to contact someone.

2) VMS speeds the flow of communication, it has the ability to eliminate typing required for company memos or letters. With VMS verbal memos can be sent to individuals with less chance of error, misunderstanding, or misinformation.

3) VMS can incorporate security measures by means of a password.

4) VMS eliminates interruptions. telephone calls can be source of irritation if they interrupt important meetings or vital work sessions.

5) VMS solves time zone problems. Corporations dealing on a global basis find time zone differences a problem when relying on traditional telephone systems. Within VMS messages can be sent to any location and stored in the recipient's mailbox. The message is stored, and as soon as the recipient arrives, the message can be accessed.

In addition to the above technologies, the concept of Teleconferencing is also taking roots in the office automation. Much research is going on in this area. Teleconferencing is simply a meeting, using electronic means, between two or more individuals, each of whom is at a different physical location. In addition to save travel time and expenses, other benefits include improved productivity, quicker solutions to the problem for which conference is convened and capability for manager to attend several meetings at diverse locations in a day. The different

types of teleconferencing are audio teleconferncing, audiographic teleconferencing and video teleconferencing.

3.6 APPLICATIONS

VOICE SYNTHESIS APPLICATION

The greatest potential for the use of speech synthesis within the office lies with the remote retrieval of electronic mail. The user may dial into his own mailbox, the synthesis system will convert the electronic text to voice output and effectively 'read out' his messages.

A second application is also connected with the remote retrieval of information, in this case the accessing of databases. The ability to remotely interact with centralized information through telephone by linking speech synthesizer to the host computer is a boon to increasingly mobile professionals and executives.

In the offices, when the hands and eyes are busy it may be useful to have audible warnings should there be any problems with the workstation. Such warnings are more likely to be noticed than error messages on the screen.

Similarly voice output can be used very effectively in training programs. For example, a training program could be developed where instructions are given verbally as the trainee runs through examples using the keyboard and workstation display.

VOICE RECOGNITION APPLICATION

The major use of voice recognition in the office environment is to permit speech to be mapped to keyboard functions. This becomes very

helpful for example in an application like Scratch pad spreadsheet. Here the voice commands are changed into keystrokes and given to the application program as though the input came directly from the keyboard.

Another interesting application associated with speech recognition particularly in the office, is that of speaker verification. This helps in maintaining the security and provides measure to safeguard the unauthorized access to data and systems. This technology provides a means of entering password vocally. The system identifies a person by the sound of his/her own voice.

On airports in the Baggage dispatch areas where the hands and eyes are preoccupied, Voice recognition systems have already made an impact and interestingly at Chicago airport it has brought the error rate from 40% to 1%.

VOICE ANNOTATION APPLICATION

A very good application of voice annotation facility is the completion of forms by voice, rather than via a keyboard. The user may call the layout of the form to the screen and then, effectively, dictate the necessary information into the fields.

A second application is that of dictation. Since voice annotation necessitates the digitisation and storage of voice, it is also possible to use this facility for dictation. Entering dictation may be simplified by voice-editing feature available on some systems.

Within office automation, generally, there has been a move to introduce systems capable of coping with all aspects of office information - text, data, graphics, image and voice. For this purpose, the need for mixed - media mailbox arises. Mixed media mailbox systems are capable of; for instance, transmitting a textual document containing a sketch or plan together with pertinent voice comments.

CHAPTER 4

APPLICATION OF VOICE RESPONSE SYSTEM - A RAILWAY ANNOUNCEMENT SYSTEM

4.1 INTRODUCTION

In Chapter 3, different voice technologies in office systems have been discussed. These technologies are Voice Response, Voice Recognition, Voice Annotation and Voice Messaging Systems.

Our plan was to build an experimental system to demonstrate the use of voice in office automation. We had Interactive Sound System [11] available with us. It consists of Interactive Sound board and software which records, stores and plays back digitized sound. But the problem with the system is that it does not provide the programmability. To develop our own software and to record and play back sound the system was found to be of no use. Because of capability of digitized sound output, which is natural, we wanted to use it, but could not use. Next choice was to use Ms-Windows 3.1 software (by Microsoft Inc.). Windows provides the user a user-friendly environment and permits the use of mouse and icon [5]. Ms-Windows 3.1 provides the sound recorder accessory which lets user record and play back digitized sound. To use this accessory, a sound card needs to be interfaced with computer, and sound driver for that card has to be loaded. We tried to use the Interactive sound card, but it was of no use, because Interactive does not provide the sound driver. To extend further, an attempt for obtaining a solution for this problem from Interactive Inc. was made. But because of no response from Interavtive Inc.; due to unusability of the Interactive Sound board, that project was to be suspended.

Thanks to availability of Covox Speech Thing system [24], we could develop the voice response system application. As explained in the chapter 3, the voice response is the process of generating human like voice from machine. COVOX Speech Thing system allows you to convert text into speech using Text-to-Speech synthesis. (Text-to-Speech synthesis has been explained in Chapter-2.) The voice output is very clear and bit robotic. The system has capability to concatenate different strings to form a sentence smoothly. Since the data to be spoken out by the system is needed to be stored in ASCII form, the memory requirement is also very less. More about COVOX Speech Thing System can be found in Appendix.

The next step was to search for the application. There were some applications we found, which could be developed using Speech Thing system. But our aim was to develop the system which will illustrate the genuine use of a voice response system in office systems. We had to search for the application of voice, where voice output is obtained with the keyboard input. Some of the applications we found were as follows :

(1) Retrieval of account balance of a bank,

(2) Data retrieval from data bank,containing accounting data in the company,

(3) Railway inquiry system.

Since in these system we can not find the genuine use of voice output with the input through keyboard, we decided to further search for the application.

After certain brainstorming, we decided to develop "*AN EXPERIMENTAL RAILWAY ANNOUNCEMENT SYSTEM (AT KANPUR RAILWAY STATION)*", In this system operator has to enter the data through the keyboard and system outputs the data in voice. Since the data needs to be output to the people at a remote

place far away from the system, the genuine use of voice in this system can be understood. Hence our problem was resolved.

4.2 PURPOSE OF APPLICATION

The purpose of an application is to announce the status of different trains, i.e., Train No., Train Name, Place from which it is coming, Place to which it is going, Platform No., whether the train is coming on right time or is late, if so then the modified arrival time or indefinite late or once the trains are included in the Train Array. (Train Array is explained later.)

The system simultaneously displays this information of trains on the monitor screen.

4.3 DESIGN OF APPLICATION

We had decided to build the application on DOS system, since the COVOX Speech Thing system works on DOS. This system consists of two units: The Speech Thing itself which connects to itself PC parallel port and speaker/Amplifier unit. The source code has been developed in C language using Borland C++ Compiler.

The different files used by the system are

1) The main program file PROJECT.EXE,

2) The data file RAILWAY.DAT containing the information about the trains which pass through Kanpur Railway Station,

3) The Text-to-Speech synthesizing program from COVOX Speech Thing, TALK.EXE.

4) The ouput file SPEAKOUT.DAT which contains information about the trains to be announced, and is read by TALK.EXE to produce spoken output.

The RAILWAY.DAT file contains the database about the all trains which pass through Kanpur Railway Station. The file contains the following information about trains as per scheduled normally :

Train No.,

Train Name,

Place from which the train is coming,

Place to which the train is going,

Arrival time of the train,

Departure time of the train,

Platform No. on which the train comes.

The database structure used for storing information is relational in nature. The information is stored in the form of table. Each record which contains the information about a train is represented by a row of the table and each column represents a different field.

WORKING OF THE SYSTEM

The system maintains the Train Array. Train Array is simply an array of records, where each record contains the Train No., Train Name, the current status of the train i.e., Platform No., Arrival Time, Late Time, along with other information as Place from which train is coming, Place to which train is going. To make an announcement about the train, it should be included in this Train Array. While speaking out, this data is used to form a string (announcement) which is passed to file SPEAKOUT.DAT. SPEAKOUT.DAT is read by TALK.EXE program to obtain spoken messages.

The system works in two modes :

1) DATA ENTRY MODE,

2) OUTPUT MODE.

(Working of the system is shown as a flow chart in fig. 4.1)

DATA ENTRY MODE :

This mode allows the user to include the train in the Train Array. When you add the train, the data about that train is fetched from RAILWAY.DAT file and is stored in the Train Array. Hence it is the normal data (i.e., as data of trains as per normally scheduled). If the train included in the Train Array is before/late, or if train is indefinite late or canceled and/or if the train is arriving on different platform, the information about that train in Train Array has to be modified. Data Entry mode allows the user to do it. Similarly once the train is arrived, the train has to be deleted from the Train Array, so that, train will not be announced again.

Once the data entry is over one can enter into output mode.

OUTPUT MODE :

In the output mode the train announcement is started. The trains in the Train Array are sorted according to their arrival time and train information is used to form the message of announcement. This message is then passed to the SPEAKOUT.DAT file.

The train which is coming at the earliest is announced first. The train is announced thrice and then the announcement about the next trainstarts.

While the announcement is going on, the information about the trains in the Train Array is displayed on the monitor screen.

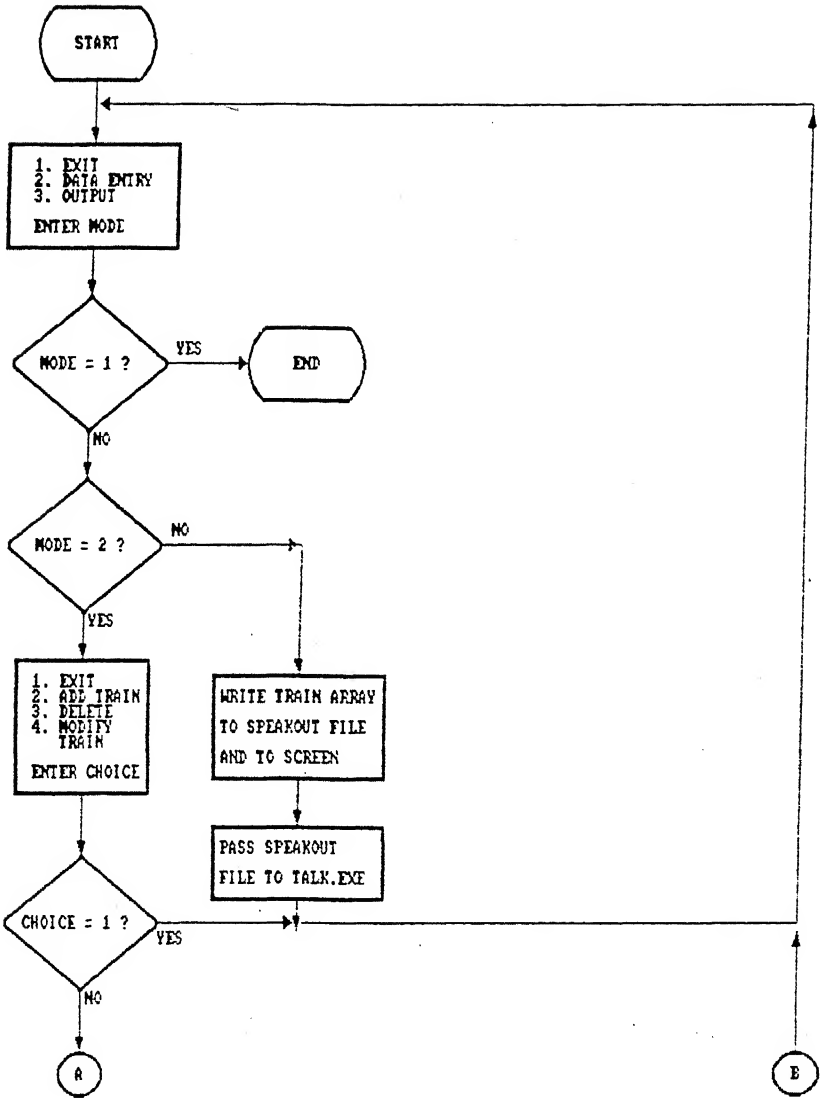


Figure 4.1 Continued...

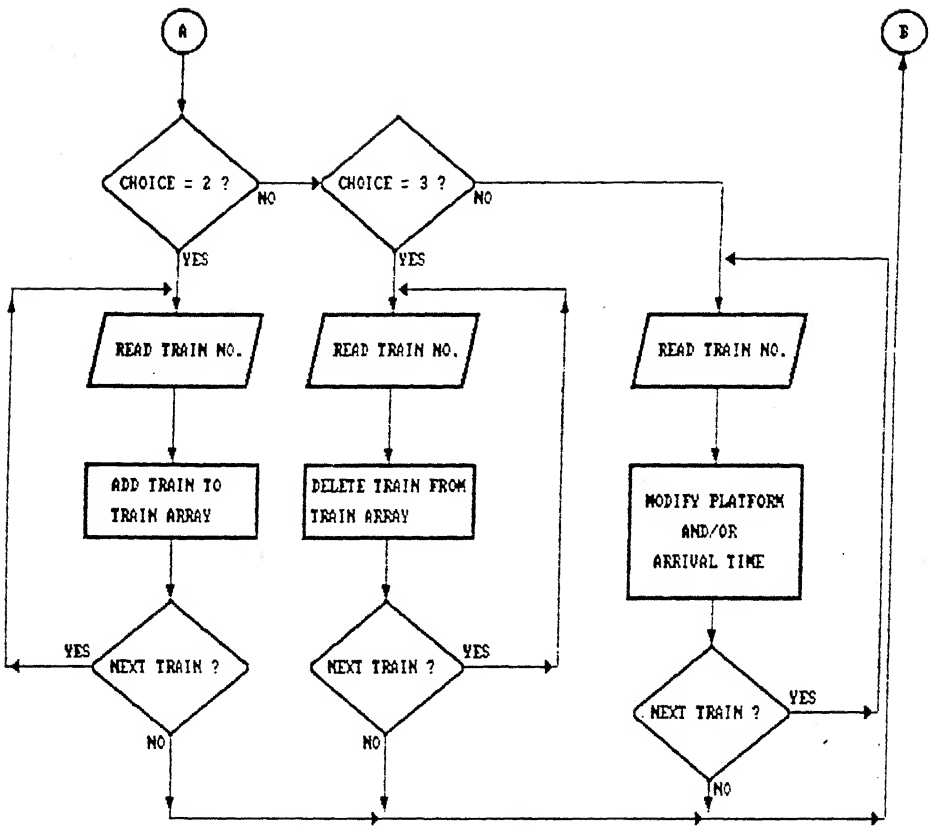


Figure 4.1 Flow diagram of Application

Once the announcement about all trains is over the system waits and asks for user to enter into the Data Entry Mode or Output Mode again. In the output mode, announcing can be interrupted by pressing space bar, so that one can delete the train from the Train Array or modify the information in the Train Array when there is a change in Platform No. or Arrival time.

4.4 SYSTEM FROM USER'S VIEW

Once the system is started, the Menu as shown in figure 4.2 appears on the screen. Entering 1, makes you exit out of the system. After entering 2, you enter into the Data Entry Mode, and Menu as shown in the figure 4.3 appears. After entering 3, the train announcement is started.

In the Data Entry Mode, entering 1 exits Data Entry Mode. Entering 2, allows the trains to be added to the Train Array. The system asks user for the Train No.. Entering 3, allows the train to be deleted from the Train Array. The system waits and asks for Train No. Entering 4, allows the user to modify information about trains which are already included in the Train Array. You can modify the Platform No. on which the train is coming or/and the train is late/before, or the train is canceled or indefinite late.For using the system, a detailed procedure is given below.

Enter 2, when menu in fig. 4.2 appears. When menu as shown in fig.4.3appears on the screen, enter 2. The system will ask for the train number. Enter the valid train no.. If the train does not pass through Kanpur Railway Station or if train no. is not valid then the system will check for it and give the error message. Also if the train is already added, the system will inform you. Once the train is added, the data about that train is fetched into the Train Array from RAILWAY.DAT file and the system asks you for addition of more trains. Once the addition is

MODE

1. EXIT
2. DATA ENTRY
3. OUTPUT

ENTER MODE :

Figure 4.2 Main Menu

ENTER 1 TO MOVE TO PREVIOUS MENU

---

2. ADD (Train about which announcement has to be made)
3. DELETE (Train which is already arrived)
4. MODIFY (Train which is already included)

ENTER OPTION :

Figure 4.3 Submenu for Data Entry Mode

completed you can enter 4, to modify the information about the train already added. When 4 is entered the system asks for train no., checks for validity of train no. when the train is entered. Against Platform press enter if you do not want change it, otherwise enter valid platform no., similarly against Late/Early press enter if you want to accept the current values of late/early hours of train, otherwise enter late/early hours. If you wantto delete any train enter 3 when menu in fig. 4.3 appears, Systems asks for the train no. to be deleted, checks for valid train no. and deletes it from the Train Array.

To make the announcement, enter 1 when menu in fig. 4.3 appears. Then menu in fig. 4.2 will appear. Enter 3 to go in output mode. System starts announcements. It announces each train thrice and starts announcing next coming train. After all trains have been announced, the systems waits and asks for entering into data entry mode or output mode again.

4.5 BENEFITS OF THE APPLICATION

The system has been designed, keeping in view the human factors. Because we felt that the designers of any system must not loose the sight of human factor. They must design the systems that people will find easy to use, else these systems will be rejected.

For the operator's benefit, the application has been designed so that it is easy to use. The system is Menu driven and self explainable. The operator has only to add trains and delete once the trains is arrived and modify information when the need arises. Since the announcement is made by the system, the operator is saved from the troublesome labor of repeating the same information again and again. Also since the system does not require the constant attention, the operator can be given another job.

Listeners of the announcement can find it very useful. Since the system announces about the same train thrice, even if the listener misses to listen once, he/she can come to know about the status of the train.

Another benefit for listener is that, the same type of voice (i.e., same pitch and speed) output is obtained always. hence once the listener becomes habitual about this voice, he/she will understand the announcement easily. In human announcement system, the disadvantage is that there is can be change in the mood of announcer like sometimes he is willing to do job, sometimes he is irritated, sometimes he feels tired; hence the voice output widely varies and may not be so clear. This disadvantage is overcome by the system developed.

Since our basic goal was to demonstrate the voice response system only, we have not gone for fool-proof system. Other facilities can be also added in the system like adding/deleting/modifying the information for the new train introduced, deleting the train automatically from the Train Array after the train leaves the station, adding more announcements for when train is leaving the station, when train is standing on the platform.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 CONCLUSION

In this thesis an application of voice response system, "A Railway Announcement System at Kanpur Railway Station" has been developed to demonstrate voice in office automation. Benefits of this application over existing human announcer system are explained. Also we have made an attempt to discuss the use of voice in office automation. Different technologies which can be used are discussed. Also the advantages and disadvantages of voice are mentioned. The advantages of voice out weigh the disadvantages.

5.2 RECOMMENDATIONS FOR FURTHER WORK

1) Digitized voice can be used in the application, so that the announcement voice is natural, hence the system will be more effective. Since the digitized voice takes a lot of memory, it is recommended that the database of trains is stored in text (ASCII) only and the other part of message is stored as a digitized voice.

2) The system can be modified as an announcement system as well as an inquiry system by interfacing phone to computer. An inquirer will dial into system, will then dial for the train whose position he wants to know. The system will speak message, which he/she can listen using his/her phone.

3) The system can be built as an announcement system at an Airport to announce the status of differnt flights.

REFERENCES

1. Anderson Eric C., Stephen Sheapard, Phil Sohn, "Signal Computing", Byte , pp 155-164, (November, 1992).
2. Bajaj K. K., "Office Automation" , Macmillan India Limited (1989).
3. Barcomb David, "Office Automation", Digital Press (1989).
4. Doswell a., "Office Automation", John Wiley & Sons, (1983).
5. Cowert Robert, "Mastering Windows 3.0", BPB Publication, (1991).
6. Cummiskey P., Jayant N. S., Flanagan J. L., "Adaptive Quantization in Differential PCM coding of speech", Bell Systems technical Journal, pp 1105-1118, (September 1973).
7. Flanagan J. L., Coker C. H., Rabiner L. R., Schafer R. W., Umeda N., "Synthetic Voices For Computers", IEEE Spectrum, pp 22-43, (October 1970).
8. Flanagan J. L., "Speech Analysis, Synthesis and Perception", Academic press Inc, (1965).
9. Hammer M., Zisman M. D., "design and Implementation of Office Information System",Office Automation : Invited Papers Infotech State-of-the-Art Report, Series 8, No.3, (1980).
10. Hirchheim R. A., "Office Automation : A Social and Organizational Perspective", John Willey and Sons, (1985).
11. Interactve Sound User's manual, Interactive Inc., USA. (1991).
12. Jarret D., "The Electronic Office", Gower Publising Co. Ltd. (1982).
13. Jauhari B. S. "Office Automation", CSI Communications, (July 1991).
14. Johnsonbough Richard and Kalin Martin, "Application Programming In C", Macmillan Publishing Company, (1990).
15. Kelley and Bootle, "Mastering Turbo C", BPB Publications (1988).

16. Lee D. L. and Lochovosky F. H., "Voice Response Systems", ACM Computing Survey, pp 351-374, (1983).

17. Markel J. D. and Gray A. H. (Jr.), "Linear Prediction of Speech", Springer - Verlag, (1976).

18. Nelson Mark, "The Data Compression Book", Prentice Hall Publication,(1991).

19. Newton S. J., "Voice In Office systems", NCC Publications, (1985).

20. Nurlin M. C. and Sprague R. H., "Information Systems Management In Practice", Prentice Hall, (1989).

21. O' Malley Michael H., "Text-To-Speech Conversion Technology", Computer , (August 1990).

22. Price S. G., "Introducing Electronic office", The National Computing Centre Ltd, (1979).

23. Rosen Arnold, "Telecommunications", HBJ Publishers, (1987).

24. Speech Thing User's Manual, Covox Inc. (May 1991).

25. Straub Detmar W. and Wethrbe James, "Information Technologies For The 1990s : An Organizational Impact Perspective", Communications of The ACM, Vol. 32, No. 11, pp 1328-1338, (November 1989).

26 Strensrud Bill and Steve Milne, "Voice Applications gaining in Office Automation", Data Management , (August 1983).

27 Witten I. H., "Principles of Computer Speech", Academic Press, (1982).

28. Ziv J. and Lempel A., "A universal Algorithm for Sequential Data compression", IEEE Transaction on Information Theory, Vol. 23, No. 3, pp 337-343, (May 1977).

APPENDIX

COVOX SPPECH THING SYSTEM

Speech Thing consists of two hardware units : The Speech Thing itself which connects to PC parallel printer port, and the Speaker/Amplifier unit. The speaker/amplifier unit may be powered either by a standard 9 V battery or by using adapter. The Speech Thing software is actually three disks full of programs which have been compressed to just two disks. Before running any of these programs, they must be decompressed. To install the software in the hard disk, insert the disk #1 in drive a: and run the INSTALL program.

Speech Thing can produce digitized sound and synthetic speech.

To produce digitized sound you need to have already recorded sound files, using COVOX Voice Master System. (extension of these files is .VMF) Using SAY program or LOADPLAY one can play back the digitized sound files. Using different arguments one can change rate of playback, enhance the higher frequencies, replace silence with noise for more natural speech.

TALK.EXE program work uses synthetic speech. It can either read an ASCII files or read individual lines typed on the screen. To read individual lines from the screen run TALK program from dos prompt. Then type a sentence and prss enter. In a few moments, you will hear the sentence spoken. To exit simply press enter on a new line. To read ASCII text files, say file named "myfile", from DOS prompt type TALK myfile. One can change the pitch and speed by giving the options.